emg2speech: synthesizing speech from electromyography using self-supervised speech models

Harshavardhana T. Gowda*
University of California, Davis
Davis, CA 95616
tgharshavardhana@gmail.com

Lee M. Miller
University of California, Davis
Davis, CA 95616
leemiller@ucdavis.edu

Abstract

We present a neuromuscular speech interface that translates electromyographic (EMG) signals collected from orofacial muscles during speech articulation directly into audio. We show that self-supervised speech (SS) representations exhibit a strong linear relationship with the electrical power of muscle action potentials: SS features can be linearly mapped to EMG power with a correlation of $r = 0.85$. Moreover, EMG power vectors corresponding to different articulatory gestures form structured and separable clusters in feature space. This relationship: SS FEATURES $\xrightarrow{\texttt{linear mapping}}$ EMG POWER $\xrightarrow{\texttt{gesture-specific clustering}}$ ARTICULATORY MOVEMENTS, highlights that SS models implicitly encode articulatory mechanisms. Leveraging this property, we directly map EMG signals to SS feature space and synthesize speech, enabling end-to-end EMG-to-speech generation without explicit articulatory models and vocoder training[2].

1 Introduction

Neural and neuromuscular interfaces hold significant promise for augmenting human abilities to interact and communicate with the external world. Brain–computer interfaces (BCIs), such as the speech neuroprostheses described in [1], [2], and [3], have demonstrated that individuals with conditions such as anarthria or amyotrophic lateral sclerosis can regain functional speech through invasive neural recordings. While such invasive approaches are well suited for individuals with severe paralysis or complete loss of articulatory control, their widespread deployment is limited by the need for surgical implantation, high cost, and clinical risk. In contrast, we propose a non-invasive speech interface that leverages preserved articulatory muscle activity, enabling a broader range of individuals—including those with laryngectomy, dysarthria, or dysphonia—to regain functional speech without the need for surgical intervention.

In this article, we present a method for leveraging self-supervised speech (SS) models to convert electromyographic (EMG) signals collected during speech articulation directly into audio, without the need for explicitly training a vocoder. Our key insight arises from the observation that speech features derived from SS models can be linearly mapped to the electrical power of muscle action potentials. Because these action potential powers corresponding to different articulatory gestures form structured and separable clusters in feature space, it follows that SS models implicitly encode articulatory information. This relationship suggests that EMG power can serve as an effective intermediate representation for mapping muscle activity to speech features. We exploit this property to design a lightweight and interpretable EMG-to-audio conversion model that leverages EMG power representations in conjunction with SS models. Such an approach has the potential to enable efficient

*Corresponding author

[2]Data and code will be made publicly available upon completion of the project.

Work in progress.

few-shot and zero-shot learning of EMG-to-audio mappings—an especially valuable property given the limited availability of EMG datasets and the frequent data distributional drift caused by factors such as electrode displacement, changes in skin moisture, and other recording variabilities.

2 Prior work

Converting non-speech signals into audio has been explored in several modalities, including lip movements-to-speech [4, 5], motor cortex neural signals-to-speech [1, 2, 6], and EMG-to-speech [7, 8]. Most existing approaches in these domains [4, 5, 1, 7, 8] assume that the alignment between the input signals (e.g., video or neural activity) and the corresponding audio is known. In contrast, we address a more challenging scenario similar to [2, 6], where the alignment between the neural activity (in our case, EMG) and speech is *unknown*. This setting requires the model not only to learn the mapping between EMG activity and audio but also to infer the underlying alignment from an exponential search space.

Work in [6, 2] addresses this alignment-free setting by training an encoder that takes motor cortex neural signals as input and learns to map them to discrete HuBERT units [9], which are then passed to a pretrained vocoder (Tacotron [10]) following the pipeline in [9]. We adopt a similar high-level pipeline for EMG-to-speech conversion. However, our approach explicitly leverages the *geometric structure* of EMG signals and their relationship to self-supervised (SS) speech representations to design an efficient encoder.

Prior work also faces several practical limitations. For instance, [6, 2] use a small-vocabulary corpus containing only 1,024 words, and in [6], each test sentence was exposed to the model an average of 6.94 times during training. Moreover, the speaking rates in these studies are restricted to 45–78 words per minute—well below the typical conversational range of 110–160 words per minute. These constraints reflect the current scope and practical limitations of existing neural speech interface systems.

To address these shortcomings, we create and open-source a large-vocabulary corpus (approximately 9 hours of EMG speech data) comprising over 6,800 unique words, articulated at a natural speaking rate of around 115 words per minute. Since data scarcity and signal distribution shifts remain core challenges in neural interface research, our approach focuses on understanding the intrinsic structure of EMG signals to guide encoder design grounded in articulatory mechanisms.

A substantial body of prior work [11, 12, 13, 14, 15, 16, 6] has laid the groundwork for the development of EMG-based speech interfaces. However, several shortcomings remain, including the use of private datasets, lack of reproducible benchmarks, and opaque architectures. Since EMG-to-speech conversion typically involves multiple components in an end-to-end pipeline, opaque designs make it particularly difficult to reproduce results and compare methods fairly. Moreover, evaluations are often limited to small-vocabulary settings (e.g., fewer than 500 words) or scenarios where alignments between EMG and audio are known a priori. In our work, we address all of these limitations.

3 Data

We collect EMG signals from 31 sites on the neck, chin, jaw, cheek, and lips using monopolar electrodes. An ACTICHAMP PLUS amplifier and associated active electrodes from BRAIN VISION (Brain Vision) are used to record EMG signals at 5000 Hertz. To ensure proper contact between the skin surface and electrodes, we use SUPERVISC, a high-viscosity electrolyte gel from EASYCAP (Easycap). We develop a software suite in a PYTHON environment to provide visual cues to subjects and to collate and store timestamped data. For time synchronization, we use lab streaming layer (LSL). See figure 1 for electrode placement. Besides 31 data electrodes, we also have a GROUND electrode (marked as GND) and a REFERENCE electrode (marked as 32). GROUND electrode is placed on the left earlobe and the REFERENCE electrode is placed on the right earlobe.

Before signal acquisition, participants were briefed on the experimental protocol and seated comfortably in a chair. Participants were instructed to articulate speech naturally. The start and end of the sentence are timestamped using mouse clicks from the subject. When a subject is ready to articulate a sentence, they click the mouse, prompting the sentence to appear on the screen. Once articulation is

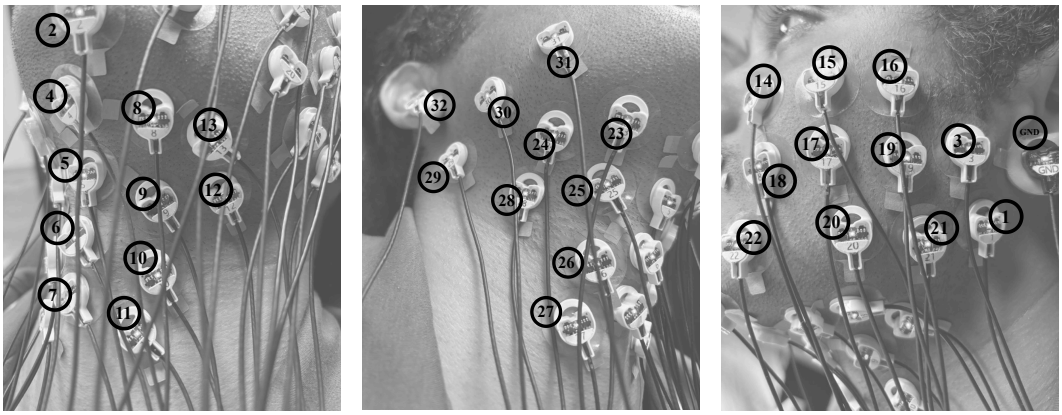

Figure 1: LEFT: Electrode placement on the left side of the neck. MIDDLE: Electrode placement on the right side of the neck. RIGHT: Electrode placement on the left cheek.

complete, they click the mouse again to indicate the end, causing the sentence to disappear from the screen—thus allowing them to articulate at their own pace.

We adapt the language corpora from [3], who demonstrated a speech brain-computer interface by translating neural spikes from the motor cortex into speech. The dataset comprises an extensive English corpus containing approximately 6,800 unique words and 9660 sentences. The corpus includes sentences of varying lengths, with the subject articulating at a normal speaking rate, averaging 115 words per minute. The dataset is divided into training, validation, and test sets containing 7000, 1000, and 1660 sentences, respectively. Sentences in the test set are not included in either the training or validation sets.

The data collection environment was carefully controlled to eliminate AC electrical interference. EMG signals underwent minimal preprocessing. The signal from the REFERENCE channel (electrode 32) was subtracted from all other EMG data channels. The resulting signals were then bandpass filtered using a third-order Butterworth filter between 80 and 1000 Hz and segmented according to sentence start and end times based on synchronized timestamps.

4 Methods

4.1 Electromyography (EMG)

EMG signals are collected by a set of sensors $\mathcal{V}$ and represented as functions of time t. A sequence of EMG signals E corresponding to articulated speech, associated with an audio signal A and phonemic content L, is represented as $E = \{\mathbf{f}_v(t)\}_{\forall\, v \in \mathcal{V}}$. Here, $\mathbf{f}_v(t)$ denotes the EMG signal captured at sensor node v as a function of time. The audio signal A encodes both phonemic (lexical) content and expressive aspects of speech such as volume, pitch, prosody, and intonation, while L represents only the phonemic content—a sequence of phonemes. For example, the phonemic content L of the word <FRIDAY> is denoted by the phoneme sequence <F-R-IY-D-AY>.

EMG covariance matrices: for an EMG signal $E_{\mathcal{V}\times\tau}$ collected from $\mathcal{V}$ sensor nodes over a duration of τ samples, we construct a symmetric positive definite (SPD) covariance matrix $\mathcal{E}_{\mathcal{V}\times\mathcal{V}} = \epsilon E E^\top$, where ϵ is a scaling factor. We denote the diagonal of $\mathcal{E}$ as $\mathbb{D}(\mathcal{E})$ and its lower triangular part as $\lfloor\mathcal{E}\rfloor$. The vector $\mathbb{D}(\mathcal{E})$ represents the muscle action potential power at each electrode $\mathcal{V}$ during the interval τ, while the off-diagonal elements capture the pairwise cross-channel covariance, reflecting the spatial co-activation structure across electrodes. A vectorized representation of $\mathcal{E}$ is denoted as $\texttt{vec}(\mathcal{E})$, a column vector of dimension $\mathcal{V}^2$.

The geodesic distance between two SPD matrices $\mathcal{E}_1$ and $\mathcal{E}_2$ is the same as the distance between their corresponding Cholesky matrices $\mathcal{L}_1$ and $\mathcal{L}_2$ and is calculated as

$$d(\mathcal{L}_1, \mathcal{L}_2) = \big\{ ||\lfloor\mathcal{L}_1\rfloor - \lfloor\mathcal{L}_2\rfloor||_F^2 + ||\log\mathbb{D}(\mathcal{L}_1) - \log\mathbb{D}(\mathcal{L}_2)||_F^2 \big\}^{1/2}, \quad (1)$$

where $||\cdot||_F$ denotes the Frobenius norm. Here, $\mathcal{L}_1$ and $\mathcal{L}_2$ are the *Cholesky factors* of the SPD matrices $\mathcal{E}_1$ and $\mathcal{E}_2$, i.e., lower triangular matrices such that $\mathcal{E} = \mathcal{L}\mathcal{L}^\top$.

EMG spectrograms: for an EMG signal $E_{\mathcal{V}\times\tau}$ collected from $\mathcal{V}$ sensor nodes at a sampling frequency f_s, we compute the short-time Fourier transform (STFT) over successive time windows to obtain a power spectrogram representation $\mathcal{S}_{\mathcal{V}\times F\times\tau'} = \left|\mathrm{STFT}(E_{\mathcal{V}\times\tau})\right|^2$, where F denotes the number of frequency bins and τ' the number of time frames. Each slice $\mathcal{S}^{(t)}_{\mathcal{V}\times F}$ captures the frequency-domain energy distribution of EMG activity across $\mathcal{V}$ electrodes at time frame t. To reduce the spectral granularity, we bin the frequency axis into B frequency bands using $\mathcal{B}_{\mathcal{V}\times B\times\tau'}(b) = \frac{1}{|F_b|}\sum_{f\in F_b}\mathcal{S}_{\mathcal{V}\times f\times\tau'}$, where F_b is the set of frequency bins assigned to band b. The matrix $\mathcal{B}^{(t)}_{\mathcal{V}\times B}$ thus represents the band power of muscle activity at each electrode across frequency bands during frame t. In practice, we use either five log-spaced bands $B_1 = [80, 125]$ Hz, $B_2 = [125, 250]$ Hz, $B_3 = [250, 375]$ Hz, $B_4 = [375, 687.5]$ Hz, and $B_5 = [687.5, 1000]$ Hz, following [17], or 31 linearly spaced frequency bands between 80 and 1000 Hz. A vectorized representation of $\mathcal{B}$ is denoted as $\mathtt{vec}(\mathcal{B})$, a column vector of dimension $\mathcal{V}B$.

4.2 Audio (A)

Audio spectrograms: for a speech waveform $a(t)$ sampled at frequency f_s, we compute a mel-scaled power spectrogram using a Hann-windowed short-time Fourier transform (STFT), followed by projection onto a mel filterbank with B mel bands. Specifically, we first obtain the power spectrogram $\mathcal{M}_{F\times\tau'} = \left|\mathrm{STFT}(a(t))\right|^2$, where F denotes the number of frequency bins and τ' the number of time frames. This spectrogram is then projected onto a mel filterbank W_{mel} spanning the frequency range $[f_{\min}, f_{\max}]$, yielding

$$\mathcal{A}_{B\times\tau'}(b,t) = \sum_f W_{\mathrm{mel}}(b,f)\,\mathcal{M}_{f,t},$$

where $b \in \{1,\ldots,B\}$ indexes the mel bands. Each vector $\mathcal{A}^{(t)}_B$ encodes the mel-band power distribution of the speech signal at frame t, emphasizing perceptually relevant frequency regions. We use $B = 80$ mel bands, $f_{\min} = 20$ Hz, and $f_{\max} = f_s/2$. We denote the column vector of an audio spectrogram by $\mathcal{A}$ throughout the article.

Audio features from SS models: for a speech waveform $a(t)$, we extract self-supervised (SS) representations by passing the signal through a pretrained model $\mathcal{S}$, yielding $\mathcal{H} = \mathcal{S}(a(t))$. The model $\mathcal{S}$ can be instantiated as WAV2VEC 2.0 [18], HUBERT [19], or WAVLM [20]. We denote the column vector of SS audio representations by $\mathcal{H}$ throughout the article.

4.3 Sequence-to-sequence models

We construct sequences of $\mathtt{vec}(\mathcal{E})$, $\mathtt{vec}(\mathcal{B})$, $\mathcal{A}$, and $\mathcal{H}$, which are emitted every 20 ms and use a context length of 25 ms. For temporal relation modeling, we employ a time depth separable convolutional network (TDS), as described below.

We use the TDS model originally designed for EMG-based keyboard typing in [21]. The model relies exclusively on local temporal context, with a 1 s causal receptive field. To improve robustness to spatial variability in electrode activity, the architecture incorporates a *Rotation-Invariance* module consisting of a linear layer followed by a ReLU activation. This module is applied to electrode channel shifts of -1, 0, and $+1$ positions, and the resulting outputs are averaged. The concatenated outputs from the Rotation-Invariance module are then fed into the TDS network for temporal modeling.

5 Results

5.1 $\mathcal{E}$ and $\mathbb{D}(\mathcal{E})$ encode articulatory information

We collected data from 12 participants, each performing 13 distinct orofacial movements, with 10 repetitions per movement. The set of movements includes <CHEEKS – PUFF OUT>, <CHEEKS – SUCK IN>, <JAW – DROPDOWN>, <JAW – MOVE BACKWARD>, <JAW – MOVE FORWARD>, <JAW – MOVE LEFT>, <JAW – MOVE RIGHT>, <LIPS – PUCKER>, <LIPS – SMILE>, <LIPS – TUCK AS IF BLOTTING>,

<TONGUE – BACK OF LOWER TEETH>, <TONGUE – BACK OF UPPER TEETH>, and <TONGUE – ROOF OF THE MOUTH>. These movements were selected to span a broad range of articulatory gestures involved in natural speech production, encompassing mechanisms such as lip rounding, jaw positioning, and tongue placement, which are essential for producing different phonemes.

Each gesture is represented by an EMG signal matrix $E_{22\times7500}$, where 22 denotes the number of electrode channels[3]. The corresponding symmetric positive definite (SPD) covariance matrix for each gesture is denoted as $\mathcal{E}_{22\times22}$, and its diagonal $\mathbb{D}(\mathcal{E})$ is a 22-dimensional vector representing the per-channel EMG power.

The vectors $\mathbb{D}(\mathcal{E})$ corresponding to different orofacial gestures naturally form distinct clusters, as shown in figure 2. We further quantified their discriminability using the unsupervised *k*-medoids clustering algorithm [22], achieving a classification accuracy of 61.41% based on $\mathbb{D}(\mathcal{E})$ (averaged across 12 subjects). When using the full covariance matrix $\mathcal{E}$ with the geodesic distance defined in equation 1, the *k*-medoids classification accuracy increased to 73.7%, both well above the random chance level of 7.69%.

These results demonstrate that both $\mathcal{E}$ and $\mathbb{D}(\mathcal{E})$ naturally encode discriminative articulatory information. While $\mathbb{D}(\mathcal{E})$ alone is sufficient to distinguish between different orofacial movements, incorporating the full covariance structure in $\mathcal{E}$ leads to improved decoding accuracy.

Note that other widely used EMG features such as log-spectrograms [21] or rectified time-domain signals [23] cannot be directly probed to verify whether such a structured representation exists. When raw EMG signals $E_{22\times7500}$ are featurized using spectrograms or rectified signals, the temporal dimension may be reduced in granularity but is not collapsed into a single frame. In contrast, covariance-based representations aggregate the temporal information within a single frame, yielding fixed-dimensional features such as $\mathbb{D}(\mathcal{E}) \in \mathbb{R}^{22}$ or $\mathcal{E} \in \mathbb{R}^{22\times22}$. (We analyze $\mathbb{D}(\mathcal{E})$ using the standard Euclidean distance, while $\mathcal{E}$ is compared using the specialized metric defined in equation 1.) Consequently, there is no low-dimensional equivalent of $\mathcal{E}$ or $\mathbb{D}(\mathcal{E})$ when using log-spectrogram or rectified features that captures articulatory structure in the same way.

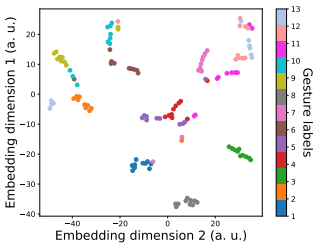

Figure 2: Different orofacial gestures are naturally separable. *t*-SNE visualization of vectors $\mathbb{D}(\mathcal{E})$ corresponding to 13 orofacial movements for a single subject. The embedding is color-coded by gesture type (*a.u.* = arbitrary units).

5.2 $\mathcal{H}$ can linearly map to $\mathbb{D}(\mathcal{E})$

We test whether there exists a linear mapping defined by a weight matrix W and bias b such that $\mathbb{D}(\mathcal{E}) \approx W\mathcal{H} + b$ with a high correlation[4].

[3]We used a subset of the 22 electrodes shown in figure 1, excluding those on the right side of the neck, for this data. Each gesture was performed over a 1.5 s interval, which corresponds to 7500 time steps at a sampling frequency of 5000 Hz. This dataset was collected independently of the one described in section 3. Unless stated otherwise, all references to data throughout the manuscript refer to the dataset described in section 3.

[4]We actually aim to probe whether $\mathcal{H}$ (768–1024 dimensions) can map to $\texttt{vec}(\mathcal{E})$ (991 dimensions). However, the resulting $\sim 10^6$-parameter linear transformation would be severely ill-posed and dominated by noise without massive data and strong regularization. To make this analysis tractable, we use $\mathbb{D}(\mathcal{E})$ as a proxy because it provides a compact, well-conditioned, and physically meaningful representation grounded in articulatory mechanisms, making it well suited for linear probing. Importantly, this substitution is justified because both $\mathcal{E}$ and $\mathbb{D}(\mathcal{E})$ encode structured articulatory information, and the latter serves as a low-dimensional surrogate for the former, as shown in section 5.1.

We use the training set described in section 3 to learn this mapping and evaluate it on the test set. We report the Pearson correlation between the predicted sequences $\mathbb{D}(\mathcal{E}')$ and the ground-truth $\mathbb{D}(\mathcal{E})$ on the test set. The representations $\mathcal{H}$ are extracted using HuBERT [19], Wav2Vec 2.0 [18], and WavLM [20]. We evaluate BASE models with a hidden dimension of 768 and 12 transformer layers, LARGE models with a hidden dimension of 1024 and 24 transformer layers, and FINE-TUNED (FT) models that have been trained for automatic speech recognition (ASR).

Correlation coefficients (r) across models and layers are shown in figure 3. We find that a simple linear model can predict $\mathbb{D}(\mathcal{E})$ from $\mathcal{H}$ with a correlation as high as $r = 0.85$. The layer-wise trends across different models partially mirror the observations reported in [24, 25] for electromagnetic articulography (EMA), where two local peaks were consistently observed across models. In our case, we observe two local peaks for Wav2Vec 2.0 models but only a single dominant peak for HuBERT and WavLM models. A sharp decline in correlation emerges in the upper layers of fine-tuned models, reflecting the growing influence of task-specific objectives. This effect is especially pronounced for Wav2Vec 2.0 compared to HuBERT and WavLM.

Notably, for the HuBERT-BASE model, the peak correlation at layer 6 aligns with the layer previously identified as optimal for discrete speech resynthesis and spoken language modeling [9]. While prior work established this empirical result, the mechanistic basis for this peak remained unclear. Our analysis provides a principled interpretation: layer 6 exhibits the strongest linear predictive power for $\mathbb{D}(\mathcal{E})$, which encodes structured and discriminative articulatory information (i.e., different articulatory gestures such as tongue and jaw positions naturally form separable clusters). This tight alignment between articulatory structure and model representations offers a direct explanation for why layer 6 is particularly effective for downstream speech resynthesis and language modeling. In short, the layer that best captures articulatory mechanisms is also the one that yields the strongest downstream performance, providing convergent evidence for its functional role.

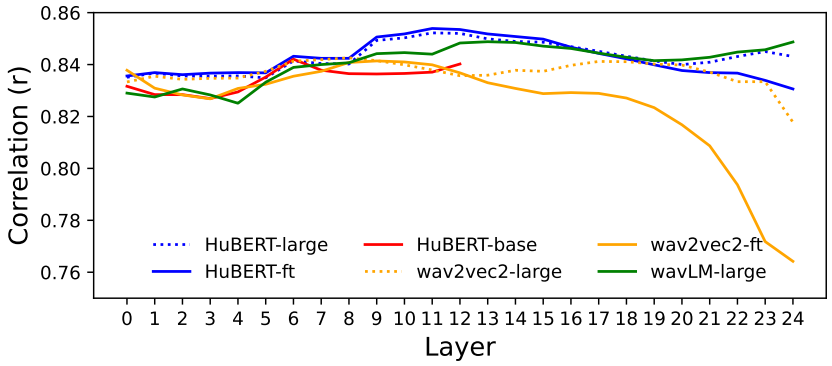

Figure 3: Layer-wise correlation (r) between $\mathbb{D}(\mathcal{E})$ and $\mathcal{H}$ across different self-supervised speech models. A simple linear mapping is used to predict $\mathbb{D}(\mathcal{E})$ from $\mathcal{H}$.

We also examined whether a similar linear mapping exists between EMG spectrogram features ($\texttt{vec}(\mathcal{B})$) and $\mathcal{H}$. Frequency bands of $\mathcal{B}$ are obtained using five log-spaced frequency bins, as described in section 4. However, the resulting correlation coefficients are substantially lower, with a maximum correlation of approximately $r = 0.57$ (figure 4). For comparison, we also computed correlations for linear mappings between $\mathcal{A}$ (audio spectrograms) and $\mathcal{B}$ ($r = 0.37$) and between $\mathcal{A}$ and $\mathbb{D}(\mathcal{E})$ ($r = 0.61$), both of which are considerably lower than the correlation between $\mathcal{H}$ and $\mathbb{D}(\mathcal{E})$.

The above observations indicate that among the different EMG feature representations considered, $\mathbb{D}(\mathcal{E})$ exhibits the strongest linear alignment with the self-supervised speech feature space $\mathcal{H}$. This strong correspondence suggests that $\mathbb{D}(\mathcal{E})$ and $\mathcal{H}$ encode highly compatible representations, making them particularly well suited for EMG-to-audio learning. In contrast, EMG spectrogram features ($\mathcal{B}$) and their alignment with audio features ($\mathcal{A}$) yield notably weaker correlations. These findings imply that, while $\mathcal{A}$ and $\mathcal{B}$ share some structure, self-supervised representations provide a more robust and articulatorily grounded intermediate latent space. Consequently, pairing $\mathbb{D}(\mathcal{E})$ with $\mathcal{H}$ offers the most effective pathway for EMG-to-audio mapping.

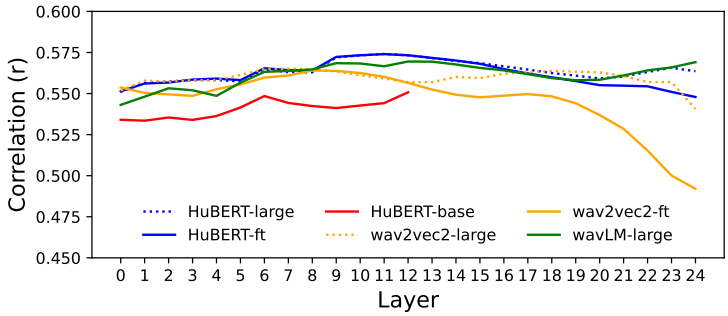

Figure 4: Layer-wise correlation (r) between $\mathcal{B}$ and $\mathcal{H}$ across different self-supervised speech models. A simple linear mapping is used to predict $\mathcal{B}$ from $\mathcal{H}$.

5.3 emg2speech synthesis

As shown earlier, the following relationship holds:

$$\mathcal{H} \xrightarrow{\texttt{linear mapping}} \mathbb{D}(\mathcal{E}) \xrightarrow{\texttt{gesture-specific clustering}} \text{OROFACIAL MOVEMENTS.}$$

The existence of a simple linear mapping from $\mathcal{H}$ to $\mathbb{D}(\mathcal{E})$ is significant: it reveals that the self-supervised representations $\mathcal{H}$ inherently encode articulatory structure reflecting underlying muscle activations. This forward direction is well posed — $\mathcal{H}$ has moderate dimensionality (768–1024), $\mathbb{D}(\mathcal{E})$ is low dimensional (31), and the mapping can be stably estimated. In contrast, the inverse problem $\mathbb{D}(\mathcal{E}) \rightarrow \mathcal{H}$ is underdetermined, non-invertible in the linear case, and especially ill-posed when temporal alignments are unknown. Nonetheless, the existence of the forward mapping provides strong evidence that $\mathcal{H}$ encodes articulatory mechanisms, motivating structured nonlinear approaches for the inverse direction rather than expecting a trivial linear inversion.

Building on this observation, we address the problem of predicting $\mathcal{H}$ from EMG features ($\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\texttt{vec}(\mathcal{B})$) without explicit temporal alignments. Since linear inversion is ill-posed, we model this mapping using a nonlinear sequence-to-sequence architecture capable of capturing the structured dependencies in $\mathcal{H}$. Specifically, EMG features are fed into a TDS convolutional network (section 4), which predicts discrete units derived from $\mathcal{H}$. We use the 100-unit discrete representation from layer 6 of the HUBERT-BASE model [9], denoted $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$. The model is trained with the connectionist temporal classification (CTC) loss [26], enabling alignment-free learning between EMG sequences and $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$. Finally, the predicted $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ sequence is passed to a pretrained Tacotron vocoder [10] to generate audio waveforms. The end-to-end architecture is shown in figure 5.

We present the results for $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ decoding in table 1. $\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\texttt{vec}(\mathcal{B})$ were provided as input to the TDS network, which was trained to predict the corresponding $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units. For example, for the sentence $_{\text{T-START}}$ <IT WAS PAID FOR>$_{\text{T-END}}$ with target $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units 71-12-71-12-4-12-4-40-93-86-13-58-32-1-99-..., the TDS model is trained to learn the mapping from $\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\texttt{vec}(\mathcal{B})$ to $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units using the CTC loss. During inference, the model outputs probabilities for all 100 $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units at each time step, and we decode these outputs using greedy search. For instance, the decoded sequence might be 71-12-57-4-54-40-93-86-13-58-16-14-76-6-36-.... We compute the unit error rate (UER) as the Levenshtein distance between the target and predicted $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ unit sequences, normalized by the length of the target sequence.

Table 1: Unit error rate (UER) for different EMG feature representations when predicting $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units. The dataset and preprocessing details are described in section 3. Lower UER is better.

MODEL INPUT	UER (% ↓)
$\texttt{vec}(\mathcal{B})$	**64.18 ± 0.68**
$\mathbb{D}(\mathcal{E})$	**62.96 ± 0.41**
$\texttt{vec}(\mathcal{E})$	**58.7 ± 0.49**

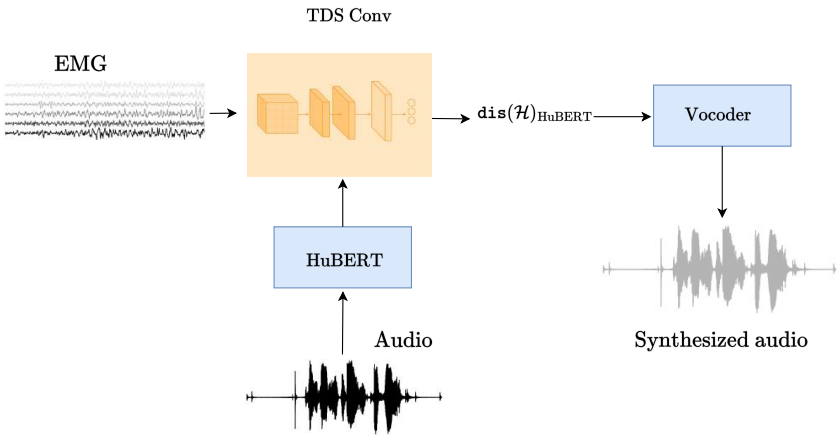

Figure 5: Multivariate EMG signals are converted into $\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\mathcal{B}$, and then passed through a TDS CONV block to predict $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$, which are subsequently fed into a vocoder to synthesize audio. Frozen neural network components are shown in blue, and trainable components are shown in orange.

We also present the results of phoneme-level decoding in table 2. For the sentence $_{\text{T-START}}$ <IT WAS PAID FOR>$_{\text{T-END}}$ with the corresponding phonemic transcription IH-T SPACE W-AA-Z SPACE P-EY-D SPACE F-AO-R, the TDS model is trained to learn the mapping from $\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\texttt{vec}(\mathcal{B})$ to phoneme sequences using the CTC loss. During inference, the model outputs probabilities for all 40 English phonemes at each time step, and the predictions are decoded using greedy search. For example, the decoded output might be IH-T SPACE W-AA-Z SPACE P-EY-T SPACE F-AO-R. We compute the phoneme error rate (PER) as the Levenshtein distance between the target and decoded phoneme sequences, normalized by the length of the target sequence.

Table 2: Phoneme error rate (PER) for different EMG feature representations when predicting phonemes. The dataset and preprocessing details are described in section 3. Lower PER is better.

MODEL INPUT	PER (% ↓)
$\texttt{vec}(\mathcal{B})$	**53.79 ± 2.02**
$\mathbb{D}(\mathcal{E})$	**50.17 ± 0.66**
$\texttt{vec}(\mathcal{E})$	**41.42 ± 0.77**

As shown in tables 1 and 2, $\texttt{vec}(\mathcal{E})$ outperforms $\texttt{vec}(\mathcal{B})$. $\mathcal{B}$ was computed using 31 linearly spaced frequency bins, and for any given time frame, both $\texttt{vec}(\mathcal{E})$ and $\texttt{vec}(\mathcal{B})$ have 991 dimensions. Notably, even $\mathbb{D}(\mathcal{E})$, which has only 31 dimensions (i.e., a dimensionality lower by roughly the square root of the others), performs better than $\texttt{vec}(\mathcal{B})$. This finding is consistent with the linear mapping results shown in figures 3 and 4.

6 Continuing work

As shown in tables 1 and 2, the model decodes phoneme sequences more accurately than $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units. We are currently exploring phoneme-guided decoding strategies for $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units to further improve accuracy. In addition, we are developing methods and models to objectively assess the perceptual quality of the synthesized audio and to compute metrics such as word error rate (WER) and character error rate (CER).

ETHICAL STATEMENT

Research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with the University of California, Davis Institutional Review Board Administration protocol 2078695-1. All participants provided written informed consent. Consent was also given for publication of the deidentified data by all participants. Participants were healthy volunteers and were selected from any gender and all ethnic and racial groups. Subjects were aged 18 or above, were able to fully understand spoken and written English, and were capable of following task instructions. Subjects had no skin conditions or wounds where electrodes were placed. Subjects were excluded if they had uncorrected vision problems or neuromotor disorders that prevented them from articulating speech. Children, adults who were unable to consent, and prisoners were not included in the experiments.

ACKNOWLEDGMENTS

This work was supported by awards to Lee M. Miller from: Accenture, through the Accenture Labs Digital Experiences group; CITRIS and the Banatao Institute at the University of California; the University of California Davis School of Medicine (Cultivating Team Science Award); the University of California Davis Academic Senate; a UC Davis Science Translation and Innovative Research (STAIR) Grant; and the Child Family Fund for the Center for Mind and Brain.

Harshavardhana T. Gowda is supported by Neuralstorm Fellowship, NSF NRT Award No. 2152260 and Ellis Fund administered by the University of California, Davis.

CONFLICT OF INTEREST

H. T. Gowda and L. M. Miller are inventors on intellectual property related to silent speech owned by the Regents of University of California, not presently licensed.

AUTHOR CONTRIBUTIONS

- Harshavardhana T. Gowda: Conceptualization, Mathematical formulation, concept development, data analysis, experiment design, data collection software design, data collection, manuscript preparation.
- Lee M. Miller: Conceptualization and manuscript preparation.

References

[1] Maitreyee Wairagkar, Nicholas S Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M Miller, Leigh R Hochberg, David M Brandman, and Sergey D Stavisky. An instantaneous voice-synthesis neuroprosthesis. *Nature*, pages 1–8, 2025.

[2] Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023.

[3] Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

[4] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional gan. *Advances in Neural Information Processing Systems*, 34:2758–2770, 2021.

[5] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13796–13805, 2020.

[6] Kaylo T Littlejohn, Cheol Jun Cho, Jessie R Liu, Alexander B Silva, Bohan Yu, Vanessa R Anderson, Cady M Kurtz-Miott, Samantha Brosler, Anshul P Kashyap, Irina P Hallinan, et al. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature neuroscience*, pages 1–11, 2025.

[7] David Gaddy and Dan Klein. Digital voicing of silent speech. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5521–5530, 2020.

[8] David Gaddy and Dan Klein. An improved model for voicing silent speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 175–181, 2021.

[9] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

[10] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. 2017.

[11] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel. Towards continuous speech recognition using surface electromyography. In *Ninth International Conference on Spoken Language Processing*, 2006.

[12] Arnav Kapur, Utkarsh Sarawgi, Eric Wadkins, Matthew Wu, Nora Hollenstein, and Pattie Maes. Non-invasive silent speech recognition in multiple sclerosis with dysphonia. In *Machine Learning for Health Workshop*, pages 25–38. PMLR, 2020.

[13] Geoffrey S Meltzner, James T Heaton, Yunbin Deng, Gianluca De Luca, Serge H Roy, and Joshua C Kline. Development of semg sensors and algorithms for silent speech recognition. *Journal of neural engineering*, 15(4):046031, 2018.

[14] Arthur R. Toth, Michael Wand, and Tanja Schultz. Synthesizing speech from electromyography using voice transformation techniques. In *Interspeech 2009*, pages 652–655, 2009.

[15] Matthias Janke and Lorenz Diener. Emg-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2375–2385, 2017.

[16] Lorenz Diener, Gerrit Felsch, Miguel Angrick, and Tanja Schultz. Session-independent array-based emg-to-speech conversion using convolutional neural networks. In *Speech Communication; 13th ITG-Symposium*, pages 1–5, 2018.

[17] Patrick Kaifosh, Thomas R. Reardon, and CTRL labs at Reality Labs. A generic non-invasive neuromotor interface for human-computer interaction. *Nature*, 645:702–711, 2025.

[18] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[19] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[20] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[21] Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean R Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alexandre Gramfort, and Michael I Mandel. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[22] Leonard Kaufman and Peter J. Rousseeuw. Partitioning around medoids (program pam). In *Wiley Series in Probability and Statistics*, pages 68–125. John Wiley & Sons, Inc., Hoboken, NJ, USA, March 8 1990. Retrieved 2021-06-13.

[23] D. M. Halliday and S. F. Farmer. On the need for rectification of surface emg. *Journal of Neurophysiology*, 103(6):3547, June 2010.

[24] Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K Anumanchipalli. Evidence of vocal tract articulation in self-supervised learning of speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[25] Cheol Jun Cho, Abdelrahman Mohamed, Alan W Black, and Gopala K Anumanchipalli. Self-supervised models of speech infer universal articulatory kinematics. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12061–12065. IEEE, 2024.

[26] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.